Interoperable genome annotation with GBOL, an extendable infrastructure for functional data mining.

Jasper J. Koehorst1,*, Jesse C.J. van Dam1,*, Jon Olav Vik2, Peter J. Schaap1 and Maria Suarez-Diez1

1. Laboratory of Systems and Synthetic Biology, Wageningen University & Research, Stippeneng 4, 6708 WE, Wageningen, The Netherlands
2. Centre for Integrative Genetics (CIGENE), Department of Animal and Aquacultural Sciences (IHA), Faculty of Life Sciences (BI-OVIT), Norwegian University of Life Sciences (NMBU),
*Authors contributed equally

Background
A standard structured format is used by public (bio) sequence databases to present genome annotations. The current format provides limited support for data provenance and data mining, hampering comparative analyses at large scale.

We have developed Genome Biology Ontology Language (GBOL) and associated infrastructure (GBOL stack). GBOL provides a consistent representation of functional genome annotations linked to the provenance.

GBOL ontology

An interoperable genome annotation pipeline

The Empusa code generator.

Empusa can be used to define an ontology and generate an associated application programming interface (API) that can be used to perform data consistency checks.

The use of Empusa ensures consistency within and between the ontology (OWL), the Shape Expressions (ShEx) describing the graph structure and the content of the resource.

Availability

- RDF2graph: van Dam et al. Journal of Biomedical Semantics 2015
  - https://github.com/jessevdam/RDF2Graph
- Empusa: bioRxiv
  - https://jgitlab.com/Empusa
- GBOL:
  - https://jgitlab.com/gbol
  - Documentation & namespace: https://gbol.life/0.1/
- SAPP Koehorst et al Bioinformatics 2018
  - https://jgitlab.com/sapp

This work has received funding from the Research Council of Norway, No. 248792 (Digisal) and from the European Union FP7 and H2020 under grant agreements No. 305340 (INFECT), No. 635336 (EmPowerPutida) and No. 634940 (MycosynVac).